ESCOLA SUPERIOR DE ENGENHARIA E GESTÃO – ESEG

TRABALHO DE ESTATÍSTICA 1 E ANALISE DE DADOS

Eduardo Felipe Sales dos Santos| RA: 39712 | Engenharia da Computação
Felipe Ultramari Domingues | RA: 39788 | Engenharia da Computação

SÃO PAULO, 10 DE JUNHO DE 2022

1) INTRODUÇÃO

Hodiernamente, é evidente que o Brasil enfrenta diversos problemas sociais, econômicos e políticos. Dentre eles, destaca-se a dificuldade de gerir a educação pública, que forma, por conseguinte, alunos com muitas defasagens nas mais diversas áreas do conhecimento.

Com o objetivo de entender melhor essa defasagem, foi escolhido o exame SARESP (Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo) da edição de 2021 como tentativa de compreender de modo mais profundo o desempenho dos alunos e suas respectivas situações. Contudo, é importante salientar que a base de dados escolhida para a análise conta apenas com o desempenho dos alunos na parte de língua portuguesa do exame.

Considerando que a base de dados é passível de diversas análises, é importante especificar o caminho que será tomado pela análise de dados. Sendo assim, a análise terá dois principais segmentos:

1- Analisar o desempenho dos alunos por si só, e, posteriormente, tentar relacionar com o fator geográfico

-Existe alguma relação geográfica com o desempenho dos alunos? Onde moram as pessoas com os melhores e piores desempenhos?

2- Comparar o desempenho dos alunos do nono ano com o terceiro ano do ensino médio, a fim de analisar a qualidade do ensino médio fornecido pela rede pública: Afinal, o aluno, ao realizar o ensino médio, mantém, melhora ou piora seu nível de conhecimento em português em relação ao nono ano?

Tal estudo é relevante e de importância significativa, visto que, ao ter essas respostas, torna-se possível entender de maneira mais clara o que acontece com os alunos, e, dessa forma, fica clara a importância de criar soluções efetivas que visem melhorar o nível de português dos estudantes da rede pública.

Ao passo que as intenções do projeto foram esclarecidas, a base de dados será apresentada, bem como toda a análise.

2) COLETA DE DADOS

Como mencionado anteriormente, será utilizada a *database* do desempenho dos participantes em relação à parte de português da prova SARESP 2021, que foi extraída da internet. Os dados foram originados pelo governo de São Paulo e disponibilizados no site do mesmo.

Link:https://dados.educacao.sp.gov.br/dataset/microdados-de-alunos-do-sistema-de-avalia%C3%A7%C3%A3o-de-rendimento-escolar-do-estado-de-s%C3%A3o-paulo

3) DICIONÁRIO DE DADOS

A base de dados, em íntegra, conta com as seguintes variáveis, sendo cada uma representada por uma coluna na planilha. Vale lembrar que as linhas são os alunos (uma linha equivale a um aluno).

Base de dados – desempenho em português no exame SARESP 2021		
VARIÁVEIS(COLUNAS)	TIPO DE VARIÁVEL	DESCRIÇÃO
SERIE/ANO	Categórica nominal	Distingue as séries dos alunos que realizaram o exame: - 5º Ano do EF; - 9º Ano do EF; - 3º Ano do EM;
PERÍODO	Categórica nominal	Período em que o aluno estuda: - Manhã - Tarde - Noite
SEXO	Categórica binária	Sexo dos alunos: - M: Masculino; - F: Feminino.
REALIZOU O EXAME (S=1, N=0)	Categórica binária	Composta por dois valores: - 1: Indica que o aluno realizou o exame; - 0: Indica que o aluno não realizou o exame.
ACERTOS (Total de questões: 24)	Quantitativa	Indica a quantidade de acertos na prova, variando de 0 a 24.
PORCENTAGEM DE ACERTOS	Quantitativa	Retorna o percentual de acerto do aluno, variando de 0% a 100%.
STATUS	Categórica ordinal	A partir do nível das questões acertadas, inclui o aluno em um grupo: -Abaixo do básico (menos de 50% de acerto) -Básico (desempenho de 50% a 69%) -Adequado (desempenho de 70% a 89%) -Avançado (desempenho de 90% a 100%)
CLASSIFICAÇÃO	Categórica ordinal	A partir da quantidade de questões acertadas, inclui o aluno em um grupo: - Insuficiente; - Suficiente; - Avançado.
REGIÃO METROPOLITANA DE SÃO PAULO	Categórica nominal	Indica a Região Metropolitana de São Paulo em que o aluno estuda. - Interior -Região Metropolitana de São Paulo -Região Metropolitana de Campinas -Região Metropolitana da Baixada Santista -Região Metropolitana de Ribeirão Preto -Região Metropolitana de Sorocaba -Região Metropolitana do Vale do Paraíba e Litoral Norte

Vale lembrar que a prova é feita pelos alunos do 3º, 5º, 7º e 9º anos do Ensino Fundamental e da 3ª série do Ensino Médio. Contudo, a base de dados contém apenas o desempenho dos alunos do 5º e 9º anos do Ensino Fundamental (EF) e o 3º ano do Ensino Médio (EM) do estado de São Paulo. É importante lembrar que as provas são proporcionais ao nível de ensino, ou seja, a prova do terceiro ano do ensino médio é diferente da do nono ano.

Conforme o dicionário de variáveis mostra, a quantidade de acertos vai de 0 a 24 questões (0-100%) e o status de 'vazio' representa os alunos que não realizaram o exame

Conhecendo as variáveis dos dados, bem como o objetivo da análise de dados, já é possível introduzi-la e extrair informações da base de dados consideravelmente grande.

4) ANÁLISE DE DADOS/ INTERPRETAÇÃO

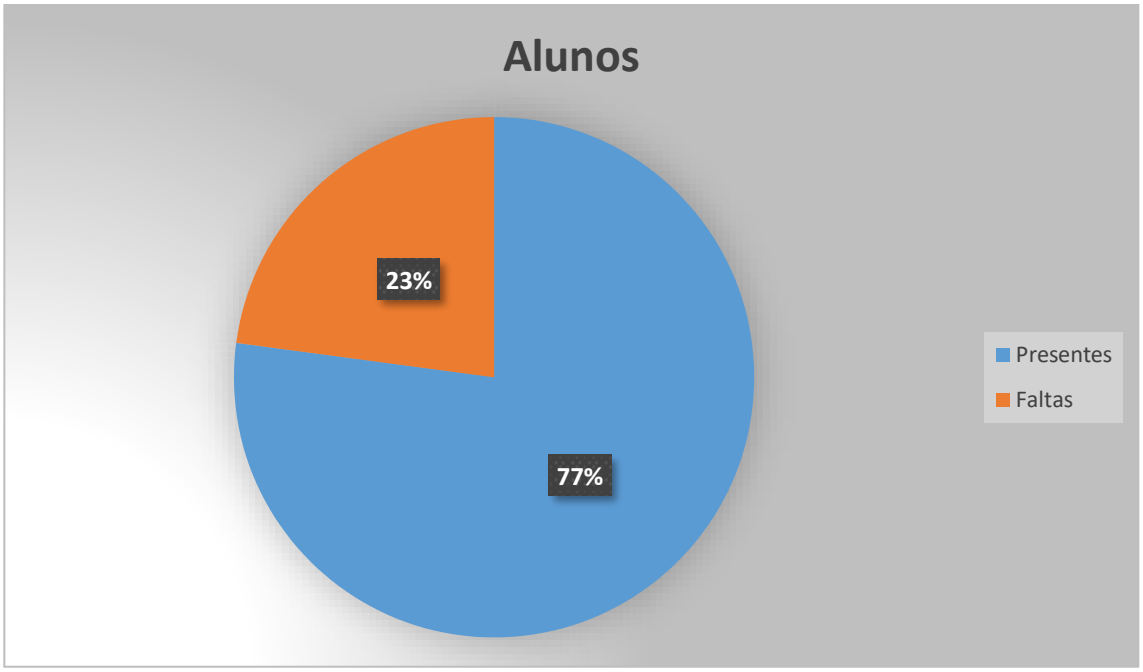
Foi possível realizar diversas análises com essa base de dados tão vasta. Para a análise, foram utilizadas as linguagens de programação python (sobretudo, as bibliotecas pandas e matplotlib, aprendidas na matéria "linguagem de programação") e R (sobretudo, com o tidyverse e ggplot) tanto para a construção dos gráficos, quanto para a filtragem e manipulação dos dados.Ambos os códigos serão disponibilizados, bem como a base de dados

Ressaltando novamente que a planilha não trabalhará com todas as turmas que fizeram o exame, mas apenas os 5º e 9º anos do Ensino Fundamental (EF) e o 3º ano do Ensino Médio (EM).

- ANALISANDO OS DADOS:

1- Quantas pessoas fizeram/não fizeram o exame?

- 457.915 alunos foram convocados para fazer o exame.
- 353.092 alunos realizaram o exame (77,11%)
-104.823 alunos não realizaram o exame (22,89%)



Analisando por região, temos que:

1º -Região Metropolitana de São Paulo: 53.107 faltas

2º -Interior: 25.977 faltas

3º -Região Metropolitana de Sorocaba: 6.182 faltas

4º -Região Metropolitana do Vale do Paraíba e Litoral Norte: 6.139 faltas

5º -Região Metropolitana da Baixada Santista: 5.523 faltas

6º -Região Metropolitana de Campinas: 3.191 faltas

Ainda que exista a possibilidade de São Paulo ter o maior número de faltas por ser a região mais populosa, um índice de 23% de faltas é considerável. Talvez campanhas de incentivo nas regiões que tiveram maior índice possam suavizar esse número.

2- ANALISANDO O DESEMPENHO POR REGIÃO:

1º) Região Metropolitana de São Paulo: 178.659 presentes – Melhor desempenho! (ou o menos pior)

A cidade de São Paulo teve desempenho médio de 58,7%, em média 14 acertos de 24 questões, com desvio padrão de 22% (5 acertos), variando entre 9 e 19 acertos.

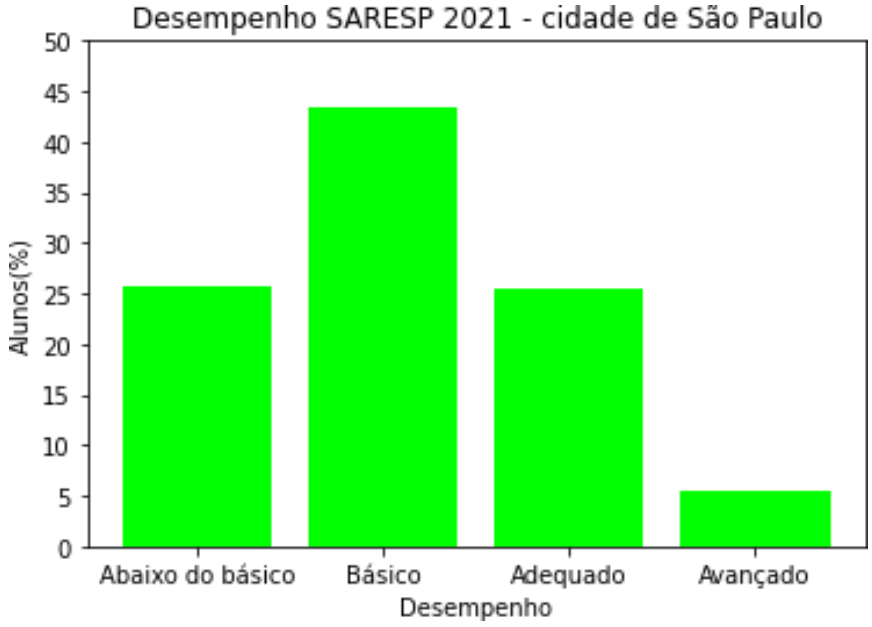
Em nível, o desempenho dos estudantes paulistanos:

25% - Abaixo do básico (45.856 estudantes).

43% - Básico (77.599 estudantes).

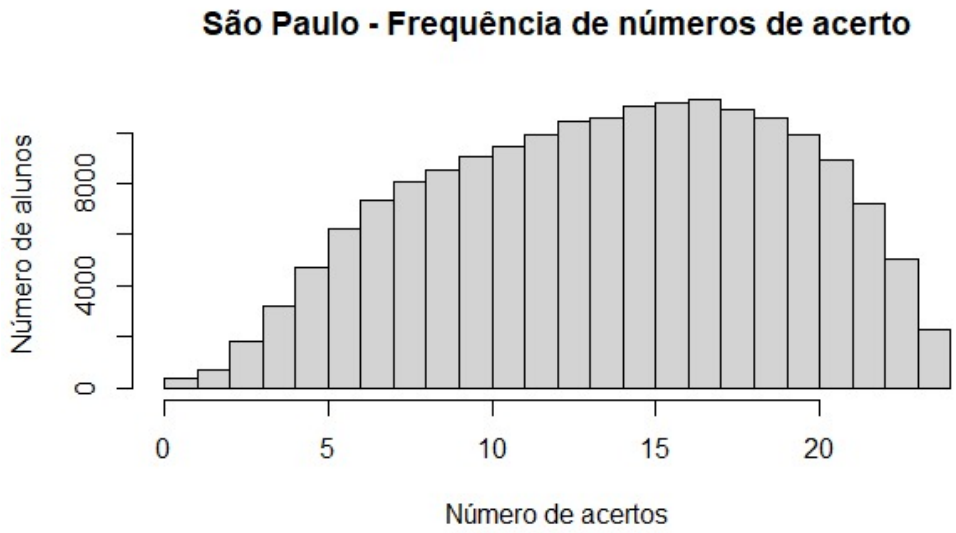
25%- Adequado (45.480 estudantes).

5,4%- Avançado (9.724 estudantes).



Os dados mostram o baixíssimo desempenho dos alunos em relação às questões de língua portuguesa. 25% dos estudantes terem um desempenho abaixo do básico e apenas 5,4% apresentar um domínio avançado da própria língua nativa revela certa defasagem do ensino. A maioria se encontra em nível básico.

Para melhor visualização do número de acertos:



O histograma acima é assimétrico à esquerda, ou seja, mais pessoas chegaram perto da nota máxima do que da nota mínima. Contudo, a existência de estudantes que zeraram a avaliação e acertaram menos de 10 questões (com certa frequência, inclusive) é uma informação preocupante. O acerto mais frequente foi de 17. É o histograma com maior frequência de números pertos de acertar todas as questões.

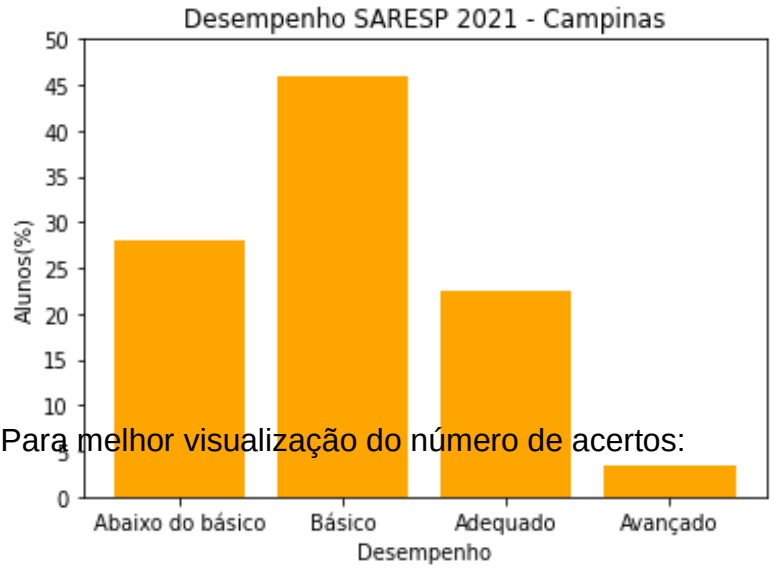
2º) Região Metropolitana de Campinas: 10.991 presentes.
Aproveitamento médio: 58,15%, em média 13,9 acertos.

25,33% - Abaixo do básico (2.784 estudantes).

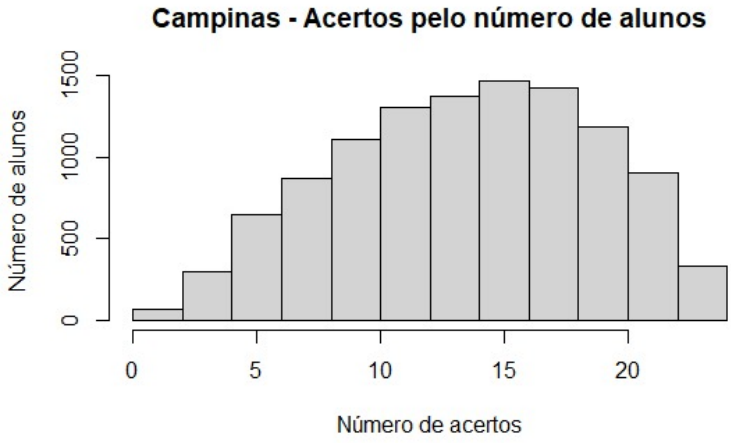
46,62% - Básico (5.124 estudantes).

24,61% - Adequado (2705 estudantes).

3,44% - Avançado (378 estudantes).



Para melhor visualização do número de acertos:



Mais uma vez, um histograma com assimetria à esquerda. Contudo, o número de acertos mais frequente está entre 16-17.

3º) Sorocaba: 19.097 presentes

Aproveitamento médio: 58,15% → em média 13,9 acertos.

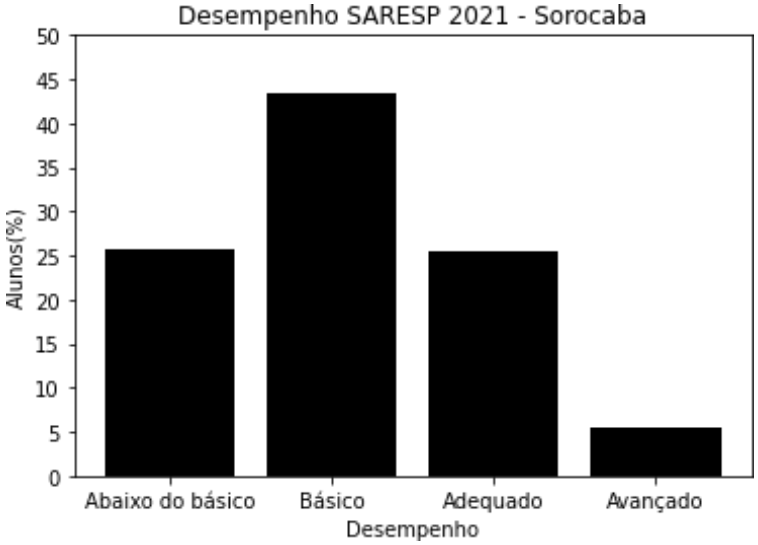
Desvio padrão: 21(%)→ 5 pontos acima ou abaixo da média.
Boa parte dos dados está entre 9 e 19 acertos.

27,40% - Abaixo do básico (5.232 estudantes).

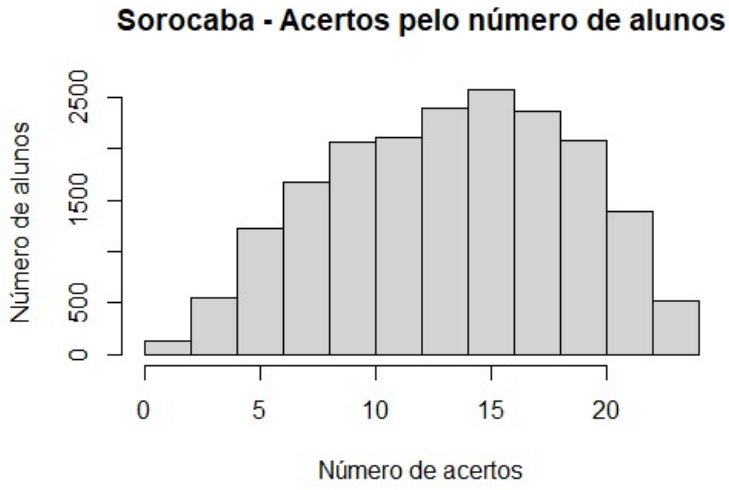
46,41% - Básico (8.863 estudantes).

23,24% - Adequado (4.439 estudantes).

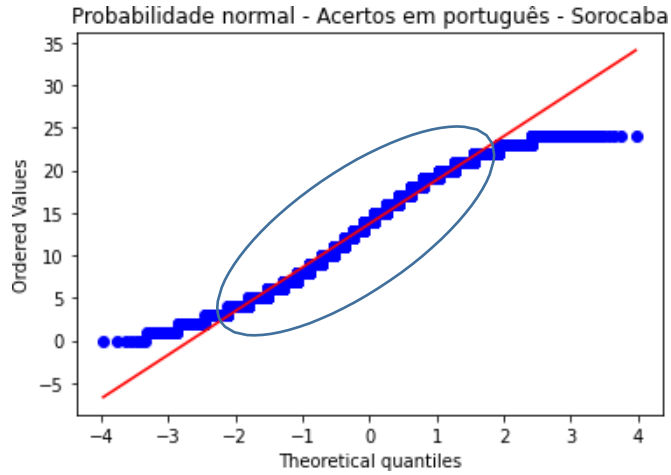
2,95% - Avançado (563 estudantes).



Para melhor visualização do número de acertos:



Levemente assimétrico pra esquerda, muito próximo de uma distribuição normal. Isso é ruim, visto que o número de estudantes que tiveram desempenho próximo de 0 acertos está quase igual ao número de estudantes que tiveram desempenho próximo de 24 acertos. O acerto mais frequente foi 15.



4º) Vale do Paraíba e Litoral Norte: 23.152 pessoas

Aproveitamento médio: 57,48% → em média 13,8 acertos.

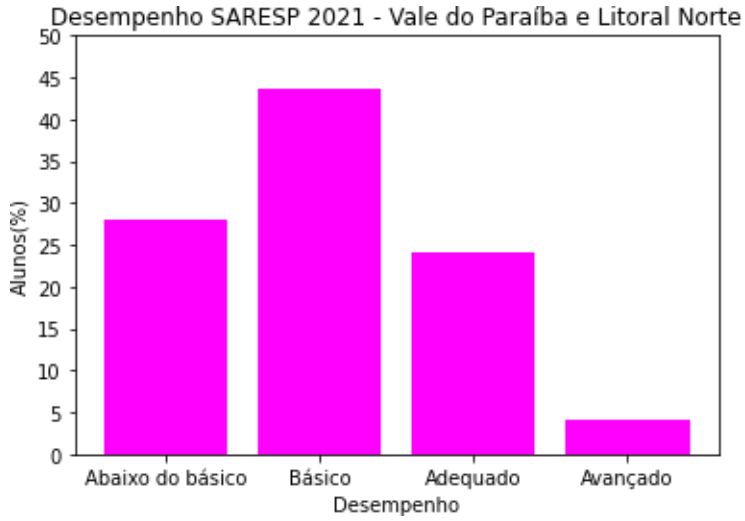
Desvio padrão: 22(%)→ 5,2 pontos acima ou abaixo da média.
Boa parte dos dados está entre 9 e 19 acertos.

27,96% - Abaixo do básico (6.474 estudantes).

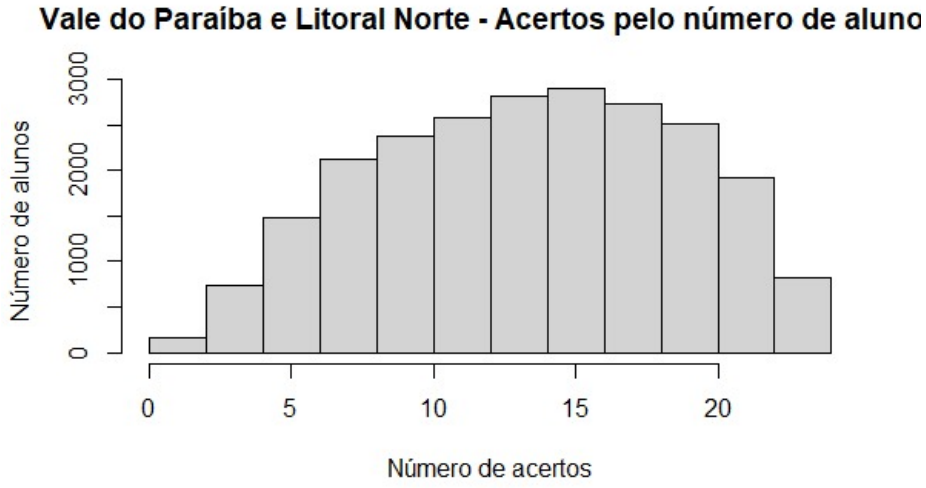
43,71% - Básico (10.119 estudantes).

24,18% - Adequado (5.597 estudantes).

4,16% - Avançado (962 estudantes).



Para melhor visualização da frequência de cada acerto:



Novamente, um histograma levemente assimétrico para esquerda. Esse histograma fortifica mais uma vez a tese de que a quantia de pessoas que acertam tudo é muito próxima das que erram tudo. O número de acertos mais frequentes foi 15.

5º) Interior: 89.234 presentes

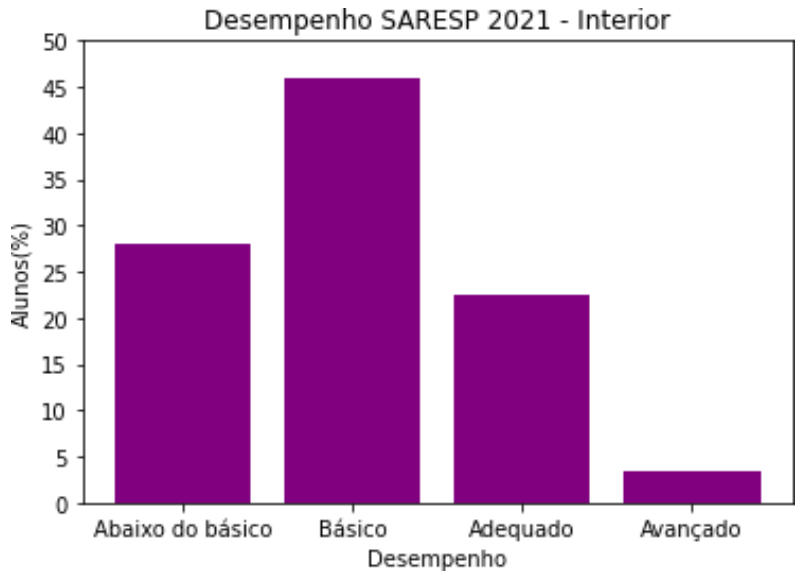
Desempenho médio: 56,6%

28,09% - Abaixo do básico (25.064 estudantes)

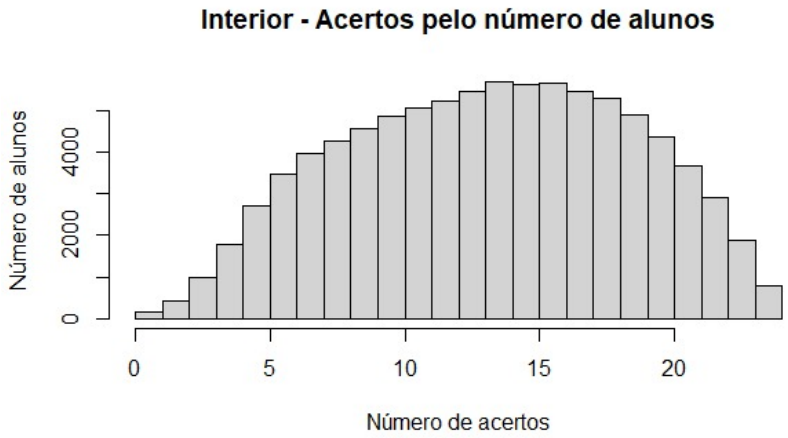
45,91% - Básico (40.097 estudantes)

22,54% - Adequado (20.019 estudantes)

3,47% - Avançado (3094 estudantes)



Para melhor visualização do número de acertos:



Esse histograma também possuí assimetria à esquerda (não tanto quanto o anterior). Ou seja, houveram mais pessoas acertando quase tudo em relação a quem não acertou quase nada. Ainda assim, é preocupante o fato de aproximadamente 3100 alunos acertarem menos de 5 questões, e cerca de 28% acertar menos da metade da prova. Idealmente, esses valores não deveriam ser frequentes. O número de acerto mais frequente foi 14.

5º) Ribeirão Preto: 15.497 presentes

Aproveitamento médio: 55,7% → em média 13,3 acertos.

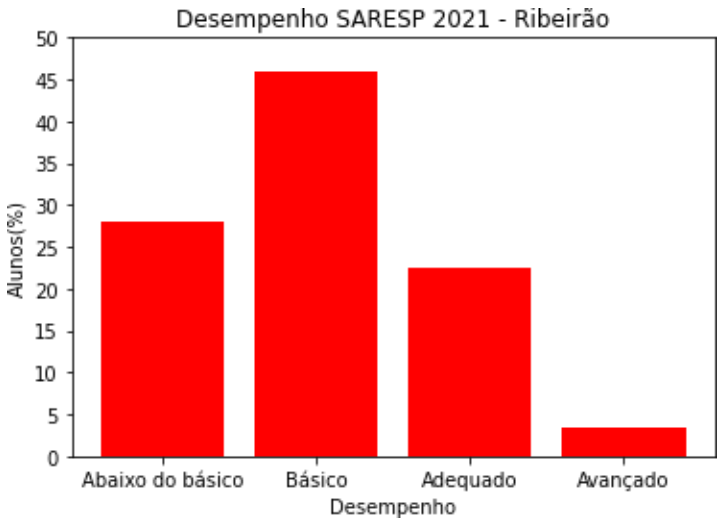
Desvio padrão: 22(%)→ 5,2 pontos acima ou abaixo da média.
Boa parte dos dados está entre 9 e 19 acertos.

30,39% - Abaixo do básico (4.710 estudantes).

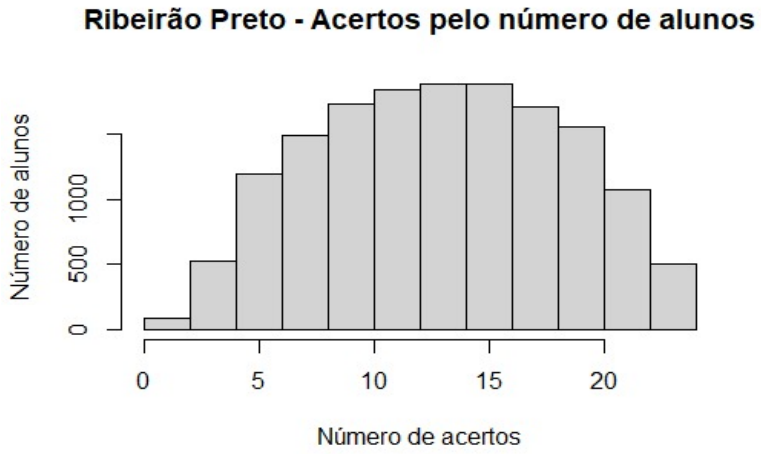
44,14% - Básico (6.841 estudantes).

21,80% - Adequado (3.379 estudantes).

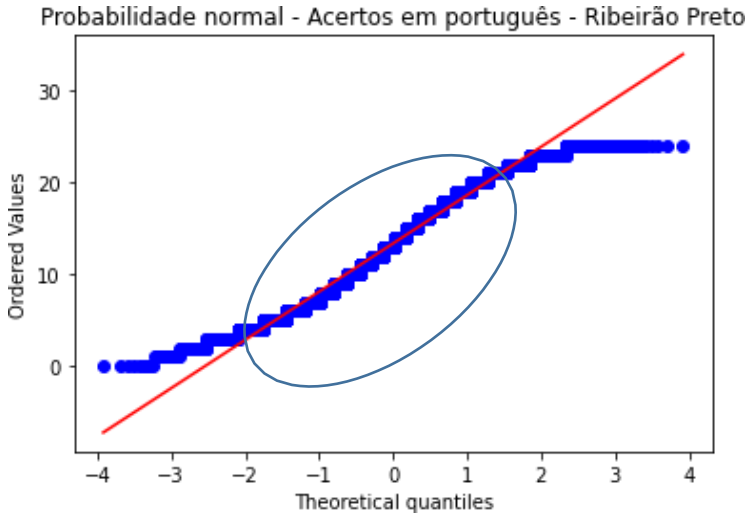
3,66% - Avançado (567 estudantes).



Para melhor visualização do número de acertos mais frequentes:



Nesse caso, o histograma segue a distribuição normal em certos intervalos, contudo, ainda é levemente assimétrico para a esquerda. O número de acertos mais frequente é de 15.



6º) Baixada Santista: 16.462 presentes: O pior desempenho

Aproveitamento médio: 54,7% → em média 13 acertos.

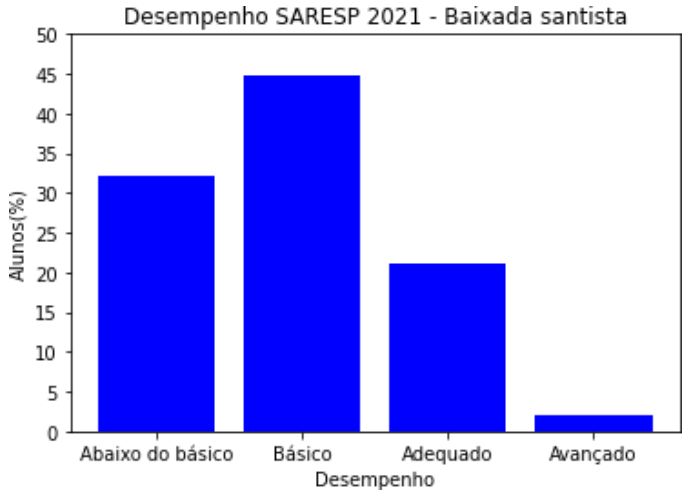
Desvio padrão: 21,7(%) → 5,2 pontos acima ou abaixo da média.
Boa parte dos dados está entre 8 e 18 acertos.

32,09% - Abaixo do básico (5.283 estudantes).

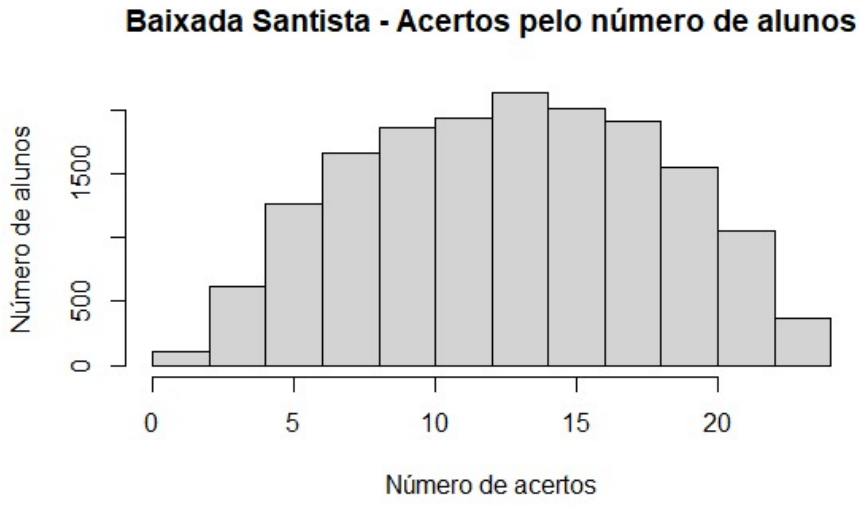
44,78% - Básico (7.371 estudantes).

21,15% - Adequado (3.482 estudantes).

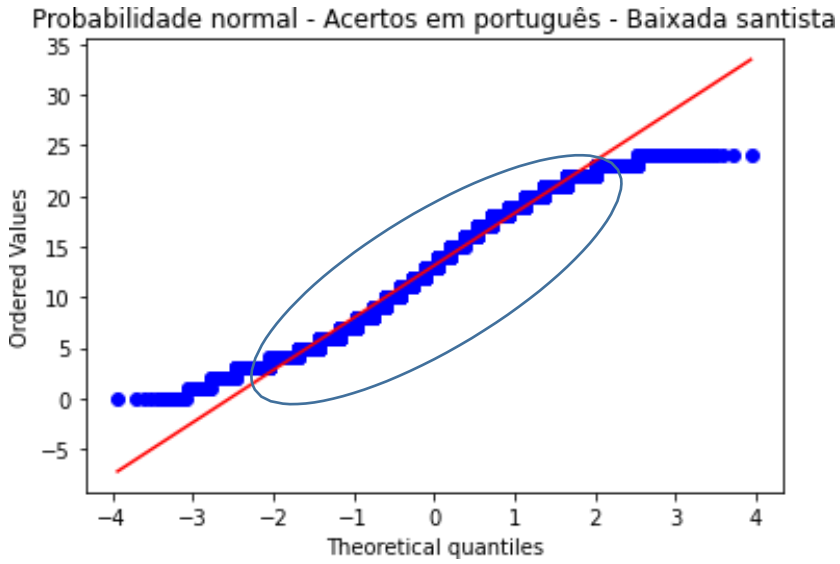
1,98% - Avançado (326 estudantes).



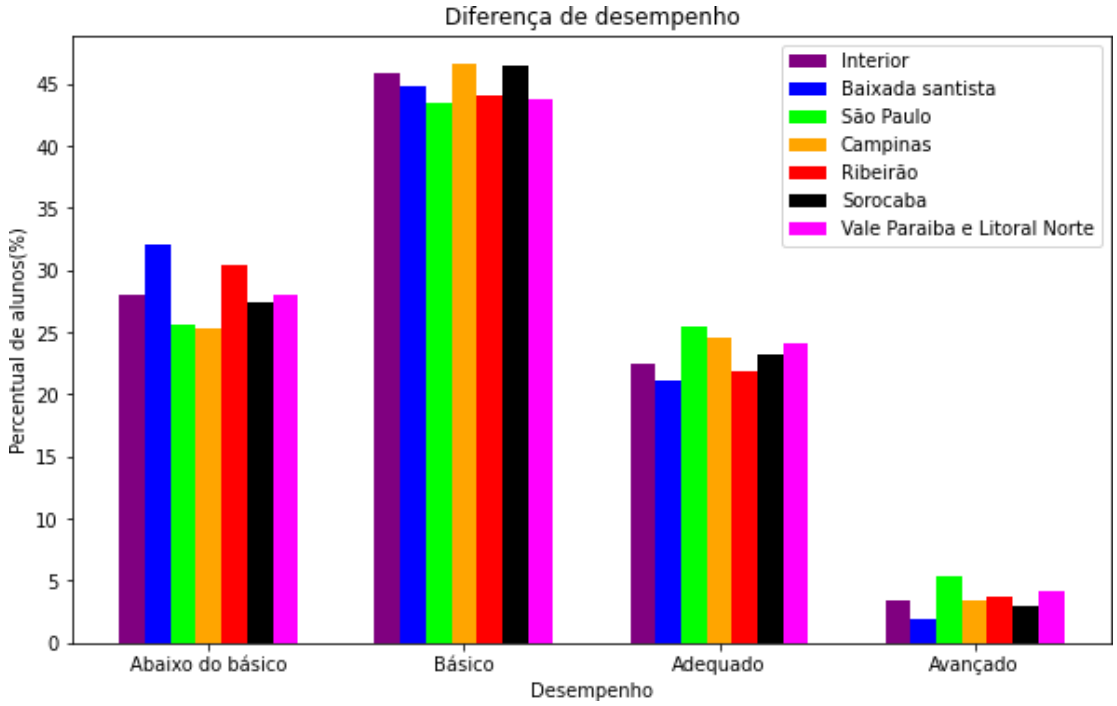
Para melhor visualização do número de acertos mais frequentes:



O número de acertos mais frequentes foi 13.Esse histograma é o que possuí o menor número de acertos mais frequentes, além de ser o mais próximo de uma distribuição normal (como anteriormente dito, nessa ocasião isso não é bom).



Aglutinando os dados, temos:



1- São Paulo: aproveitamentento médio de 58,7%, em média 14 acertos (Melhor desempenho)

2- Campinas: aproveitamento médio de 58,15%, em média 13,9 acertos

3 -Sorocaba: aproveitamento médio de 57,1%, em média 13,7 acertos

4- Vale do paraíba: Aproveitamento médio: 57,48%, em média 13,8 acertos.

5- Interior: aproveitamento médio de 56,6% , em média 13,5 acertos

6- Ribeirão Preto: aproveitamento médio de 55,7%, em média 13,4 acertos

7- Baixada Santista: aproveitamento médio de 54,7% , em média de 13 acertos (Pior desempenho)

3- Comparando o desempenho do EM-3ª série e 9º Ano EF: Qual a efetividade do Ensino médio?

É mister apontar que o quinto ano será descartado dessa parte dessa análise, uma vez que o objetivo principal é o de analisar a diferença de desempenho dos alunos antes de entrarem com o desempenho dos alunos prestes a sairem.

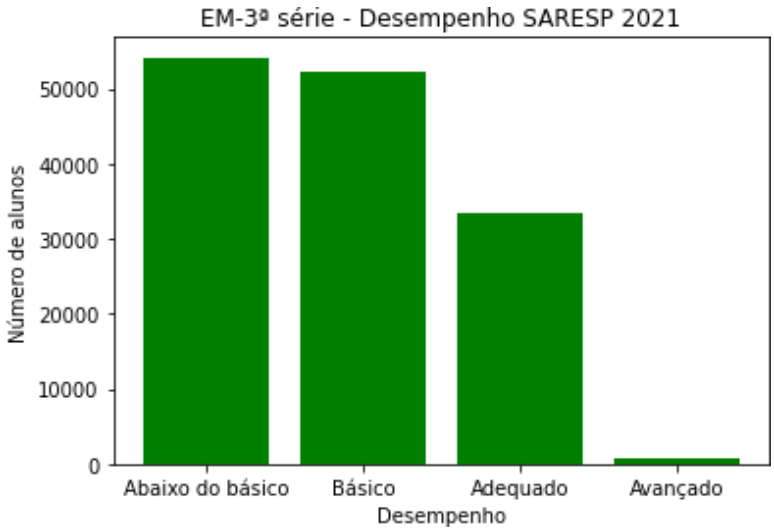
EM - 3ª série : 140.615 alunos

38,52% - Abaixo do básico (54.161 estudantes).

37,16% - Básico (52.256 estudantes).

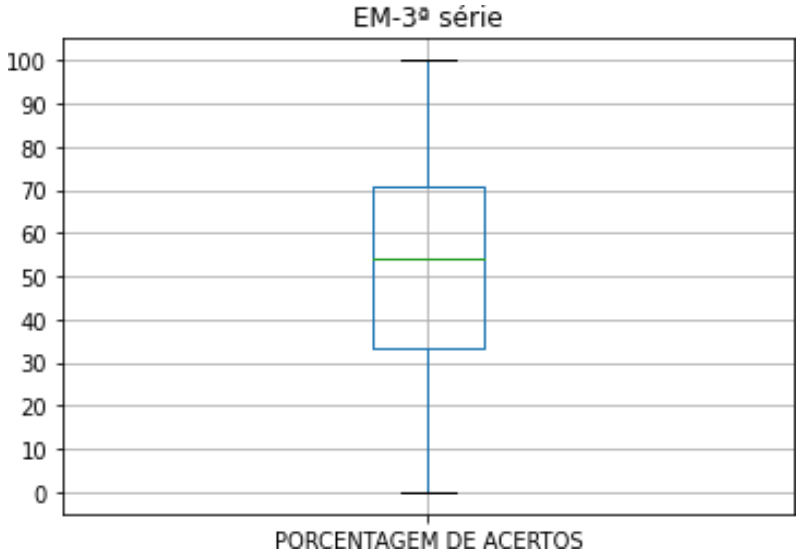
23,82% - Adequado (33.491 estudantes).

0,005% - Avançado (707 estudantes).

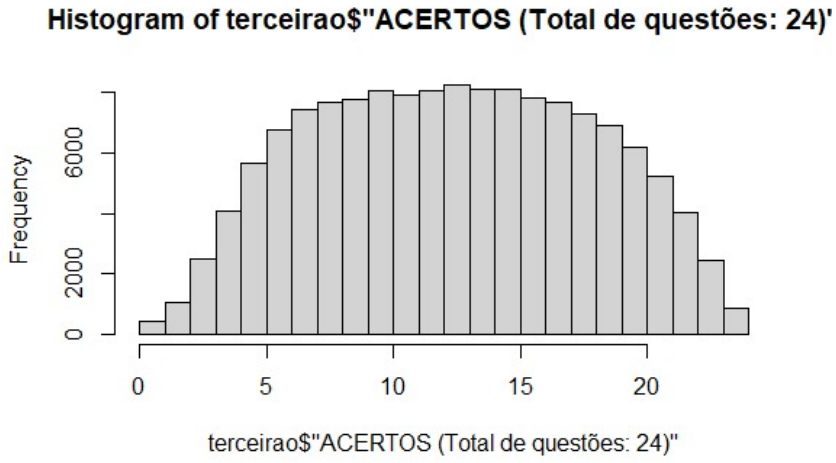


A proficiência mais obtida do terceiro ano do ensino médio foi “Abaixo do básico”, e, posteriormente, “Básico”. Tal informação revela a imensa defasagem do ensino médio público, em que os alunos, no último ano escolar, possuem majoritariamente um desempenho abaixo do básico em relação ao português.

Para melhor compreensão sobre onde os dados estão mais concentrados:



Com o *boxplot*, (os outliers não foram colocados para não poluir o gráfico),pode-se inferir que maior parte dos alunos tiveram desempenho entre 38% e 70%, com maior parte próxima de 38%.



Nesse histograma, percebe-se que muitos valores acontecem em uma frequência semelhante. A quantidade de alunos que obteram de 5 a 10 acertos é muito próxima do número de acertos mais frequente (12 acertos).

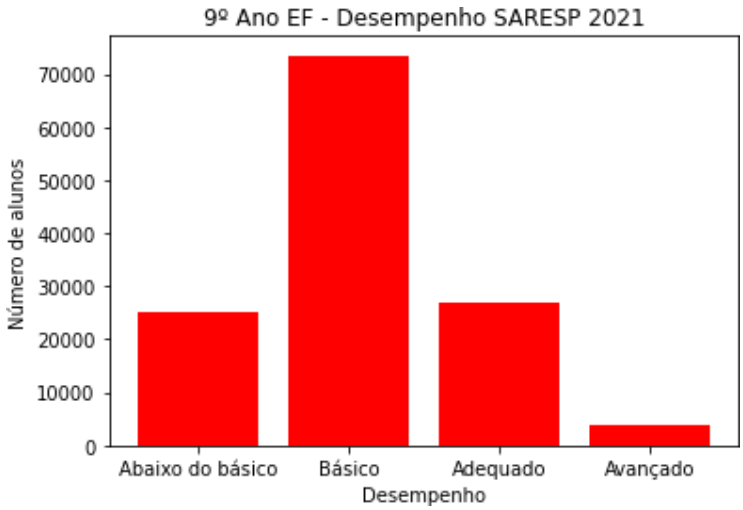
EF- 9ª série : 129.702 alunos

19,49% - Abaixo do básico (25.273 estudantes).

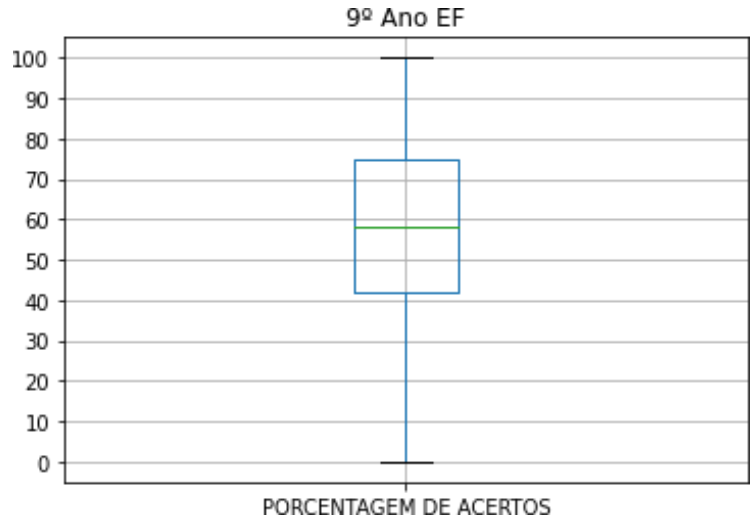
56,68% - Básico (73.509 estudantes).

20,77% - Adequado (26.933 estudantes).

3,07% - Avançado (3.987 estudantes).

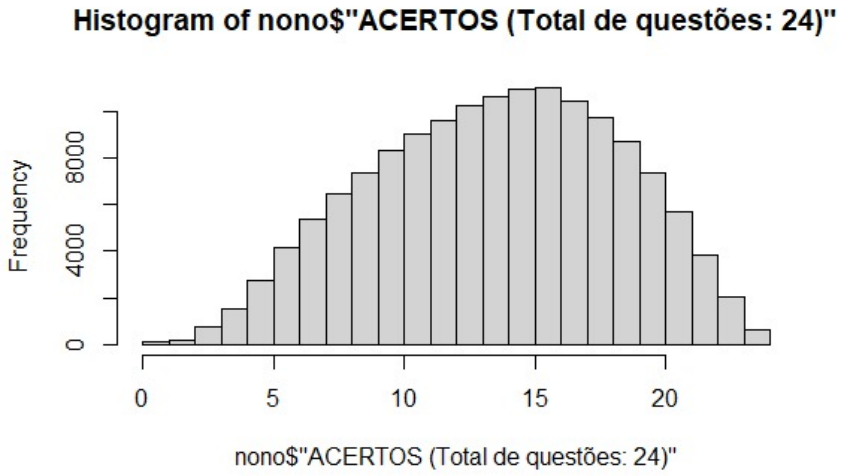


Para ver a concentração de dados:



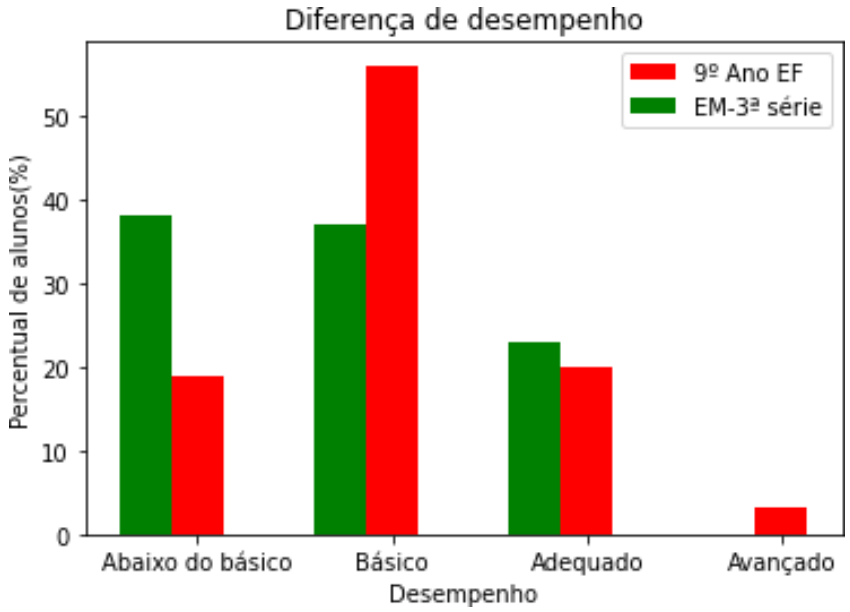
Nesse bloxpot, vemos dados mais concentrados. Enquanto no terceiro a porcentagem de acertos da maioria varia de 38% a 70% , a porcentagens de acertos do nono ano varia de 42% a aproximadamente 73%. Além de mais concentrados, eles estão com os dados concentrados em porcentagens maiores, evidenciando um melhor desempenho.

Histograma:



Por meio do histograma, notamos uma evidente assimetria à esquerda,ou seja, os acertos abaixo de 12, ainda que frequentes, não foram tão frequentes quanto os acertos acima de 12.

Aglutinando os dados, temos:



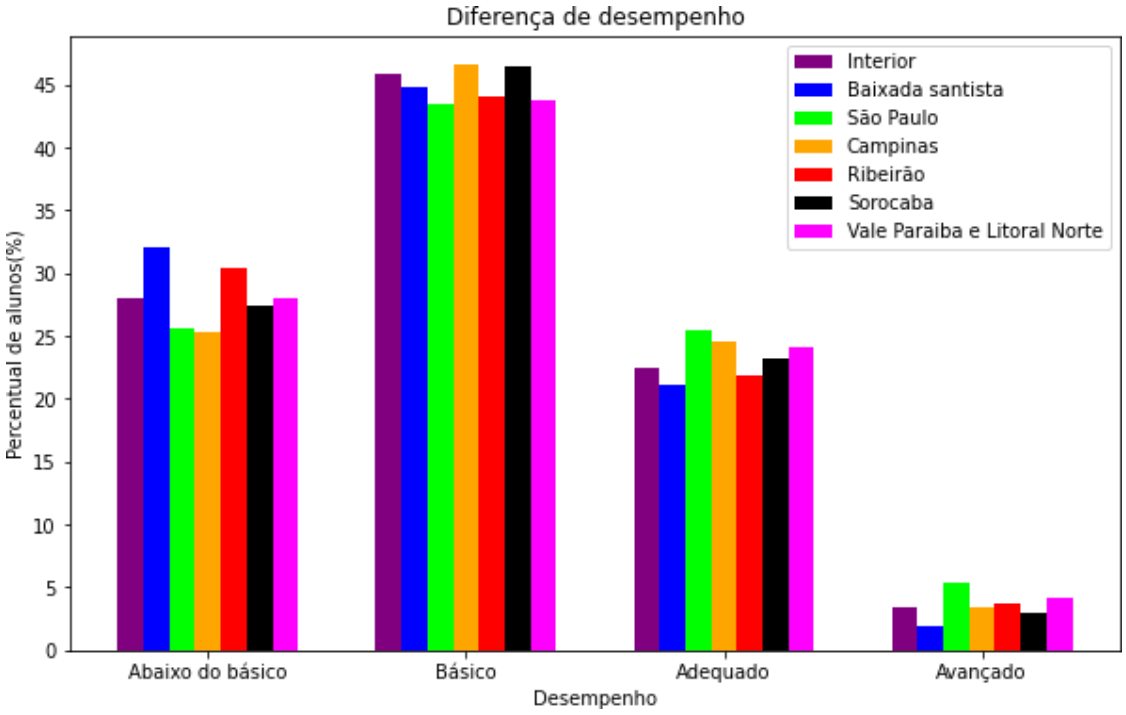
5) Interpretações finais e conclusões

1- Quantas pessoas fizeram/não fizeram o exame?

Considerando que 22,89% dos convocados não compareceram, pode ser interessante criar incentivos para que mais alunos compareçam. Um bônus para aqueles que realizarem o exame pode atrair mais estudantes. Diminuindo cada vez mais o número de faltas, é possível gerar dados e informações com cada vez menos viés.

2- Analisando o desempenho por região:

Considerando o desempenho de cada região, podemos concluir que:

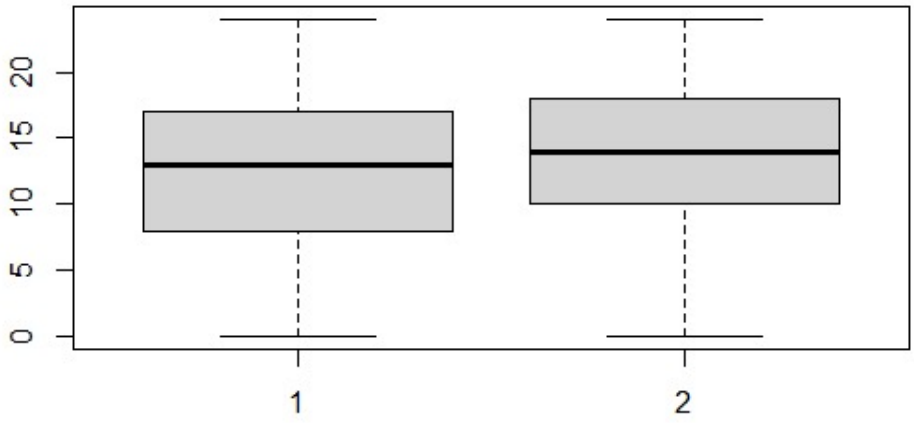


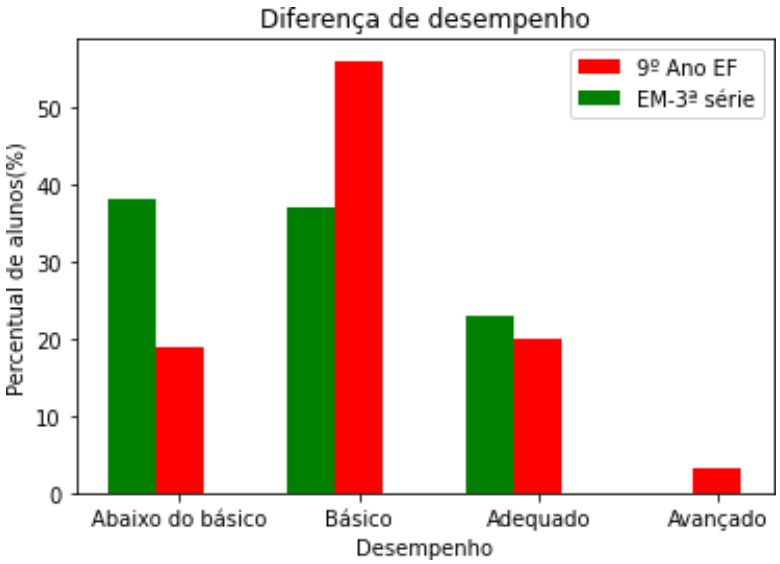
São Paulo, enquanto a região com melhor desempenho, foi a que teve o maior percentual de alunos de proficiência adequada e avançada, e também foi a que teve o menor percentual de alunos de proficiência abaixo do básico e básico. Contudo, a maioria ainda se encontra em nível básico.

Já a Baixada Santista, enquanto a região com pior desempenho, foi a que teve o menor percentual de estudantes com proficiência adequada e avançada, e também a que teve maior percentual de estudantes com nível abaixo do básico. Em síntese, a maioria dos alunos de todas as regiões do estado de São Paulo se encontram em nível básico e abaixo do básico, ficando clara a necessidade

urgente de mudanças no sistema educacional público. Maiores investimentos na educação, contratar mais professores e melhorar a infraestrutura das escolas são algumas das formas que possam suavizar a situação.

3- Comparando o desempenho do EM-3ª série e 9º Ano EF





Em síntese, pode-se dizer que o nono ano teve melhor desempenho em relação ao terceiro. Além de ter menos alunos com proficiência abaixo do básico em relação ao terceiro ano, o nono ano também teve muito mais alunos com proficiência avançada e básica. Não só isso, como também tiveram desempenho

mais concentrado, enquanto o terceiro ano teve maior variação, tendendo a acertos menores. (Considere o boxplot 1 o terceiro ano, e o bloxpot 2 a nona série).

Torna-se evidente, portanto, uma contradição sobre o ensino médio e sua efetividade. Por que os alunos que realizaram o ensino médio possuem uma proficiência menor de domínio da língua portuguesa em relação aos alunos que estão prestes a cursar? O ensino médio não deveria, ao menos, manter o nível?